MycoKeys 5: 45–63 (2012)
doi: 10.3897/mycokeys.5.4302
www.pensoft.net/journals/mycokeys

RESEARCH ARTICLE

A peer-reviewed open-access journal
MycoKeys
Launched to accelerate biodiversity research

An appraisal of megascience platforms for biodiversity information

Dagmar Triebel[1], Gregor Hagedorn[2], Gerhard Rambold[3]

1 *Bavarian Natural History Collections, IT Center, München, Germany* **2** *Julius Kühn-Institute (JKI), Federal Research Centre for Cultivated Plants, Berlin, Germany* **3** *Mycology Dept., University of Bayreuth, Bayreuth, Germany*

Corresponding author: *Dagmar Triebel* (triebel@bsm.mwn.de)

Academic editor: *T. Lumbsch* | Received 12 November 2012 | Accepted 27 December 2012 | Published 28 December 2012

Citation: Triebel D, Hagedorn G, Rambold G (2012) An appraisal of megascience platforms for biodiversity information. MycoKeys 5: 45–63. doi: 10.3897/mycokeys.5.4302

Abstract

The megascience platforms Biodiversity Heritage Library (BHL), Catalogue of Life (CoL), Encyclopedia of Life (EOL), Global Biodiversity Information Facility (GBIF), International Barcode of Life (iBOL), International Nucleotide Sequence Database Collaboration (INSDC) and JSTOR Plant Science, all belong to a group of global players that harvest, process, repurpose and provide biodiversity data on all kinds of organisms. Each of these platforms primarily focus on one data domain, for instance, taxonomy and classification, occurrence, morphology, ecology, and molecular data.

The present contribution describes aspects of processing and provision of biological research data on these platforms, focusing on the technical implementation of data exchange, copyright issues, and data sharing policies as well as their implications for data custodians, owners, providers, and publishers. With the exception of JSTOR Plant Science, most international initiatives seek long-term business models and funding mechanisms to provide online data openly and free of charge. For example, currently GBIF depends on governmental commitments for its funding, and CoL is financed by EU or national grants, as well as being based on Species 2000, a British non-for-profit company, and ITIS. These business models are compared with that of JSTOR Plant Science, the commercial portal of the Global Plant Initiative (GPI). All initiatives currently meet challenges of sustainability with regard to data curation as well as software development for maintaining the complexity of their services. All platforms discussed here also harvest and provide mycological and lichenological research data.

Copyright *Dagmar Triebel et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License 3.0 (CC-BY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Keywords
Internet Platforms for Natural Sciences, BHL, CoL, EOL, GBIF, iBOL, JSTOR Plant Science, INSDC, DDBJ, EMBL, GenBank, Barcoding, Data Flow, Research Data

Introduction

In biodiversity research, data driven approaches, relying on internet resources that provide huge amounts of quality information, are increasingly important. In the late 1990s, most biodiversity websites offered more or less static web content and were operated by individual scientists or research groups. At that time, only a limited number of data access portals, mostly addressing data collections of homogenous structure, existed. Today, web-based information sources are almost overwhelmingly complex, heterogeneous, and seemingly exponentially growing. To find useful and reliable biodiversity information, several general approaches exist: (a) web sites where individual scientists or scientific community members curate categorized link collections, e.g., The Mycology.Net (http://www.mycology.net), (b) global search providers such as Google, Bing, or Yahoo and others that provide solutions with advanced generic search tools, and (c) so-called megascience platforms which have been set up in a scientific community context. The present contribution will analyse the latter approach and the probable challenges these will have to face in the future. It will focus on seven large platforms for biodiversity, which are relevant for lichen research data at a global scale.

Some major biodiversity data projects and platforms which have a geographically limited scope such as the Atlas of Living Australia (ALA; http://www.ala.org.au/) and the envisaged European LifeWatch project (http://www.lifewatch.eu) are not subject of this paper. Some other limited time projects, e.g., EDIT (http://www.e-taxonomy.eu/) or 4D4Life (http://www.4d4life.eu/), are not discussed in detail here because their results are contributing or have contributed to other platforms (e.g., 4D4Life results are injected into CoL).

Finally, several new initiatives or platforms are under active technical development and might attract relevant amounts of biodiversity and ecology research data in the near future. They are, however, not yet suitable for a comparison of the kind intended here. ViBRANT (http://vbrant.eu/) develops web-based virtual research communities for biodiversity science. Based on Scratchpads (http://scratchpads.eu/) and the Biowikifarm (http://biowikifarm.net), individual research communities share data management, curation, analysis and publishing services. This allows to improving effectiveness of research and supports long term data preservation and re-use in several of the platforms discussed here. pro-iBiosphere (http://www.pro-ibiosphere.eu/) is a coordination project to provide for a global generic organismic knowledge publishing and curation platform that brings the traditional Flora and Fauna editorial efforts into the digital world. The Map of Life (MOL; http://www.mappinglife.org/about/) project is an initiative that is just starting. Supported by content data from GBIF and EOL, it focuses on occurrence maps along with tools for quering and transforming related data.

History and scope of megascience platforms processing biodiversity information

Starting in the early 1990s, researchers in biology recognized the importance of the internet for disseminating data for research purposes. Work groups dedicating themselves on nucleic acid sequence data were the first to initiate domain-specific data projects covering all organism groups at a global level. Three platforms, EMBL-Bank (http://www.ebi.ac.uk/embl/), GenBank (http://www.ncbi.nlm.nih.gov/genbank/), and DDBJ (http://www.ddbj.nig.ac.jp) emerged, which in 1992 formed the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org). Today, this consortium provides access to several databases focussing on molecular data.

Ten years later, in 2001, two other megascience platforms were initiated by scientists with the objective to collect and curate organismic biodiversity information. The first was the Catalogue of Life (CoL) that aims to produce a global quality-assured checklist of all species of plants, animals, fungi and other macro- and micro-organisms known to science (http://www.catalogueoflife.org). Currently, this data pool is supplied by data sets of more than 100 taxonomic databases and checklists and is annually updated. CoL currently contains authorative names and synonyms for about 8,000 lichen species obtained from the Global Species Database LIAS (Rambold 2012; http://www.catalogueoflife.org/col/details/database/id/79).

In the same year, 2001, the Global Biodiversity Information Facility (GBIF) was initiated. It provides species distribution data in the form of occurrence records along with names and classifications, as well as links to additional information (http://data.gbif.org/tutorial/tutorial). GBIF makes data from more than 400 so-called 'data publishers' from all over the world openly and freely available. Occurrence records with geographical coordinates are visualized in global distribution maps. For instance, for Lecanoromycetes 3,281,898 occurrence records exist (last visited: 31-10-2012).

In 2003, the precursor project ('API – African Plant Initiative') of the Global Plant Initiative (GPI; http://gpi.myspecies.info) was started. The output of the efforts of GPI is accessible via the JSTOR Plant Science portal providing access to foundational content concerning plant type specimen data, taxonomy, references, high-resolution images of type specimens, and related literature (http://plants.jstor.org/action/about). JSTOR Plant Science makes available data that are shared by more than 220 partner herbaria worldwide. Certain lichen type collections, like those of BM, G, H, LINN and M, are accessible as well.

Subsequently in 2005, the Biodiversity Heritage Library (BHL) consortium was founded. BHL is a consortium of libraries with a focus on natural history and botanical literature that cooperate in digitizing and making legacy literature of biodiversity accessible under open access (http://www.biodiversitylibrary.org; last visited: 26-06-2012). Currently, more than 60,000 titles and 100,000 volumes are available. Scientific organism names in the literature are recognized by means of the uBio NameBank (including lichen species names from LIAS and Index Fungorum, see http://names.ubio.org/browser/details.php?namebankID=3871575). The BHL is not the only ini-

tiative or project digitalising historical biology literature (more than 40 are listed by Kasparek 2010), but so far it is the largest one.

In 2007, the CBOL (the Consortium for the Barcode of Life) started the International Barcode of Life (iBOL http://ibol.org) initiative. The original idea is a consequence of the barcoding proposal published by Hebert et al. (2003). The initiative is devoted to the collection of DNA barcoding sequence data http://www.barcodinglife.com/ stored in the Barcode of Life Data System (BOLD). BOLD contains 156,461 taxa species with barcode sequences and a total of 1,702,485 specimens with barcode sequences, (last visited: 26-06-2012); about 1,250 of these are Lecanoromycete specimens (http://www.barcodinglife.com/index.php/Taxbrowser_Taxonpage?taxid=262560; last visited: 04-11-2012). The primary mission of iBOL is to extend the geographic and taxonomic coverage of the barcode reference library to store the resulting barcode records, to provide community access to the knowledge they represent, and to create new devices to ensure global access to this information. The work of iBOL is carried out by a research alliance spanning 25 nations with varying levels of investment and responsibilities (http://www.barcodeoflife.org/content/about/what-ibol). The overall task of the iBOL research participants is to collect and curate specimens, to extract DNA, to gather barcode data (records of group-specific DNA marker gene sequences), and to build up an informatics platform being required for storing and providing these records for species identification.

In the same year when iBOL was launched, 2007, another highly ambitious megascience initiative was launched: The Encyclopedia of Life (EOL; http://eol.org/discover), which collects and freely provides information about all species at a global scale including classifications, multimedia data, maps of occurrences. This initiative created more than 3.3 million pages: 1,079,652 pages with some amount of content, including 94,467 with considerable contents, being called 'rich pages' (http://eol.org/statistics/page_richness?date_one_set=2012-10-12&date_two_set=2012-10-31data.gbif.org).

Data domains

Each of the major biodiversity data platforms profiled here has its own scope (Table 1). Aside, each has a focus on one of the three central information segments: names and classification, occurrence, and descriptive or trait data.

Name data primarily include accepted names, synonyms, and proposed higher classification (usually reflecting a phylogenetic concept). Data from this domain may be classified as being either unequivocal (or 'objective', like the validity of a name according to the relevant nomenclatural code as well as the obligate synonymy), or equivocal ('subjective', e.g. depending on a phylogenetic concept, like the assignment of a heterotypic synonym to a currently accepted taxon name). Relevant databases for lichenology which provide taxon names as well as taxonomic concepts are LIAS names (http://liasnames.lias.net/; Triebel et al. 2010), Species Fungorum (http://www.speciesfungorum.org/), MycoBank (http://www.mycobank.org/), and, in future, the

Table 1. Contents and scopes of megascience platforms providing and processing biodiversity information

Megascience platform	Content and scope	Year of launch	Logo
International Nucleotide Sequence Databases (INSDC)	Nucleic acid sequences	1992	
Catalogue of Life (CoL)	Taxonomic checklists	2001	
Global Biodiversity Information Facility (GBIF)	Occurrences and records	2001	
JSTOR Plant Science	Type specimens, multimedia objects	2003	
Biodiversity Heritage Library (BHL)	Biodiversity literature, multimedia objects	2005	
Barcode of Life (iBOL)	DNA barcoding sequences	2007	
Encyclopedia of Life (EOL)	Knowledge data, species fact sheets, multimedia objects	2007	

evolving Chinese Portal for fungal names (http://www.fungalinfo.net/fungalname/fungalname.html). EOL, GBIF, BOLD for iBOL, and INSDC use the names and classifications from these and other name providers. Name data are also essential for the BHL site which provides access to digital images of biodiversity literature resources. BHL extracts scientific names from the digitized documents by a taxonomic name recognition algorithm and offers extended search techniques for these names. JSTOR Plant Science needs taxonomic names and information on classification to improve search tools and provide basic data on type specimens including multi-media objects important for taxonomy and systematics.

Occurrence data may be split into two major categories: collection and observation data. Collection data are correctly considered as more reliable when compared to observational records. However, for many groups of taxa, with sufficient quality con-

trol of observer expertise and combined with digital photographs or other multimedia data, the relevance of observational data has dramatically increased in recent years. The central platform for collection and occurrence data is GBIF. GBIF set up various kinds of tools and APIs to mobilise, visualize, and analyse the distribution patterns of taxa (http://tools.gbif.org), preferably with the data contents available through GBIF.

Descriptive data may be split in various specific ones, referring to a) morphological and anatomical characters and character states, b) to chemical properties (in the case of lichens, e.g. the highly diverse secondary metabolites), and c) to nucleic acid sequences, from DNA sequences of various genes (including the so-called 'barcoding genes') to full genome sequences d) to behavioural and ecological features. The central platform for descriptive data under a), b), and d) is EOL with the limitation that the descriptions of species are generated by individuals and partners with heterogeneous content data (e.g., FishBase), and do not derive from structured database contents. One major phenotypic trait database with structured descriptive data for lichen species is LIAS light (http://liaslight.lias.net), covering the morpho- and chemodiversity of about two thirds of all known lichen species (> 9,000 taxa). The most outstanding nucleic acid sequence database repository with three partners is the INSDC consortium with EMBL-Bank, NCBI-GenBank, and DDBJ.

Business models and consortial structures

In the case of the INSDC consortium, the collaborating institutions (DDBJ, EMBL-ENA, and NCBI-GenBank) have established data-sharing policies for more than twenty years. Responsibility for the quality and accuracy of the records, however, has been assigned to the submitting authors or institutions (http://www.insdc.org/policy). The three well-established partner institutions agreed to maintain a common technical core infrastructure for submission and archiving nucleic acid sequence data worldwide (Cochrane et al. 2010).

The Catalogue of Life (CoL) consortium is a cooperation of two partners being the autonomous federation of database organizations and taxonomic database custodians 'Species2000' (registered as a not-for-profit, limited by guarantee company in the UK), and ITIS, a partnership of federal agencies and other organizations from the United States, Canada, and Mexico. The CoL secretariat is currently located at University of Reading (UK) and mainly financed by grants and financial support from one of the two partners, Species2000. Data are provided by experts from 115 taxonomic databases from around the world, each responsible for a defined group of organisms (http://www.catalogueoflife.org/col/info/about). Data quality is assured by peer-review mechanisms.

The Global Biodiversity Information Facility (GBIF) is an intergovernmental organization. GBIF members or 'GBIF participants' (http://www.gbif.org/participation/being-a-part-of-gbif/) are about 60 nations (China not included) and approximately 50 international organizations. The voting participants provide financial contribution

to the GBIF secretariat, the advisory committee structure and the work program on a yearly basis (http://www.gbif.org/governance/finance/). They are responsible for the national support of the GBIF network, which is primarily a non-centralised system with national participant nodes (http://www.gbif.org/participation/). Data are provided by more than 420 mainly institutional publishers, being responsible for data quality and accuracy. GBIF is developing a decentralised network of 'biodiversity information facilities' (BIFs) established and maintained by its participants which, e.g., are countries or international organisations that have signed the GBIF Memorandum of Understanding (MoU) (http://www.gbif.org/participation/participant-nodes).

JSTOR Plant Science has been funded and spearheaded by the Andrew W. Mellon Foundation through the project 'Global Plant Initiative' (http://about.jstor.org/content/jstor-plant-science). Content partners and publishers are represented by more than 200 institutions from over 50 countries. The major goal of the initiative is to digitise herbarised type specimens (mainly plants, but also bryophytes, algae, fungi, and lichens) and provide access to images and metadata at a global scale. The digitised and quality-controlled data is published under non-exclusive license conditions by JSTOR (http://about.jstor.org/10things). JSTOR itself is a not-for-profit organization with a commercial segment being based on the income from subscriptions fees by foundations, universitary institutions, libraries and individuals for accessing the information. A considerable number of scholarship institutions have access for free, but the majority of individual scientists who are not affiliated to such institutions can use only a limited amount of the research data from JSTOR Plant Science for free.

The Biodiversity Heritage Library (BHL) is a consortium of 12 partner libraries from US and UK natural history collections, supported by grants from several foundations. Its primary funding came from the Encyclopedia of Life initiative (http://biodivlib.wikispaces.com/Funding+Sources), a close co-operation partner of this initiative. The BHL project is focussed on digitising legacy literature related to biodiversity. Since 2009, it has expanded globally, e.g. by an EU funded project with 28 institutions involved, as well as BHL nodes in China, Australia, and Brazil.

The International Barcode of Life (iBOL) initiative with its central node in Canada is funded mainly and by the Ontario government, two Canadian Foundations, and the Genome Canada association. The international research program is coordinated by a team at the University of Guelph and supports barcoding activities of the iBOL partners to a certain degree. The governance board consists of senior staff from Genome Canada, a science advisory committee, and an international scientific collaboration committee with members drawn from nations with funded barcoding projects linked to iBOL (http://ibol.org/funding-shortfall-brings-changes-at-ibol/). iBOL is structured and organized in four major nodes (Canada, China, Europe, US), several regional and national nodes, as well as partner organizations from 27 nations (http://ibol.org/about-us/partner-nations/).

The Encyclopedia of Life (EOL) is an international consortium, financially supported by 16 institutions and 6 foundations. Its contents are provided by more than 220 partner content data platforms and more than 62,000 so-called 'members'. Data

is quality-controlled by about 300 active EOL curators on a voluntary basis (http://eol.org/statistics; access 2012-10-31). The EOL executive committee provides governance and decision-making at the policy level. The senior individuals represent GBIF, BHL, foundations in the USA, and cornerstone institutions in the USA, Australia, China, Egypt, and Mexico (http://eol.org/info/3#SC).

In conclusion, only three to four of the seven initiatives have sufficient technical infrastructure backbone that can be regarded as independent from third-party grants to scientists or scientific institutions, which are INSDC, GBIF, JSTOR Plant Science, and probably EOL. For four of the seven initiatives discussed here, financing the creation of content data is not the central issue of the business model. Only JSTOR Plant Science, iBOL and BHL-US directly back this kind of activity by financial support. The remaining ones mainly rely on the motivation of volunteers and individual enthusiasts (EOL, CoL), or on national funding programs to support generation of data and its delivery (GBIF, iBOL).

Data flows, cross-linkages

Each of the seven platforms has its own profile with respect to data domains, providers and scope of contents, and user communities, but strong dependencies between the platforms (e.g. between BHL and EOL) exist. Furthermore, there is cooperation between the four platforms GBIF, iBOL, EOL and JSTOR Plant Science to visualise occurrence data and to link data from biodiversity literature. They therefore require a common name data backbone, provided by a jointly developed technical structure in the frame of a common project, the Global Names Architecture (GNA; http://www.globalnames.org/) project. For sequence data which is produced in the iBOL context, the INSDC consortium with NCBI GenBank has agreed to stand by as the general data repository and backup archive.

The cooperation and linkages between the seven megascience platforms themselves as well as between the seven initiatives and their primary data providers is assumed to be facilitated by relying on open source principles and on contents provided under creative commons or open database licences conditions or – at least – data sharing policies on a non-exclusive basis. With growing content, the data flow and cross-linkages between the seven platforms is visible (Fig. 1). In parallel, the backtracking of multimedia data with corresponding metadata, e.g., from EOL and from thematic portals like EDIT (http://search.biocase.org/edit/: this is mirroring the GBIF index database), back to the primary providers or publishers of scientific data is possible.

The data life cycle and data flow starts with data production. The megascience platforms are harvesting infrastructures which are part of a 'food chain' that starts with the primary-content producers to primary and secondary harvesters and ends up with data users, consumers and digesters. Data harvesters like GBIF and CoL, which are typically fed by research data from individual scientists and institutions, may alternatively also be supplied by primary data collecting infrastructures, e.g. by the World Regis-

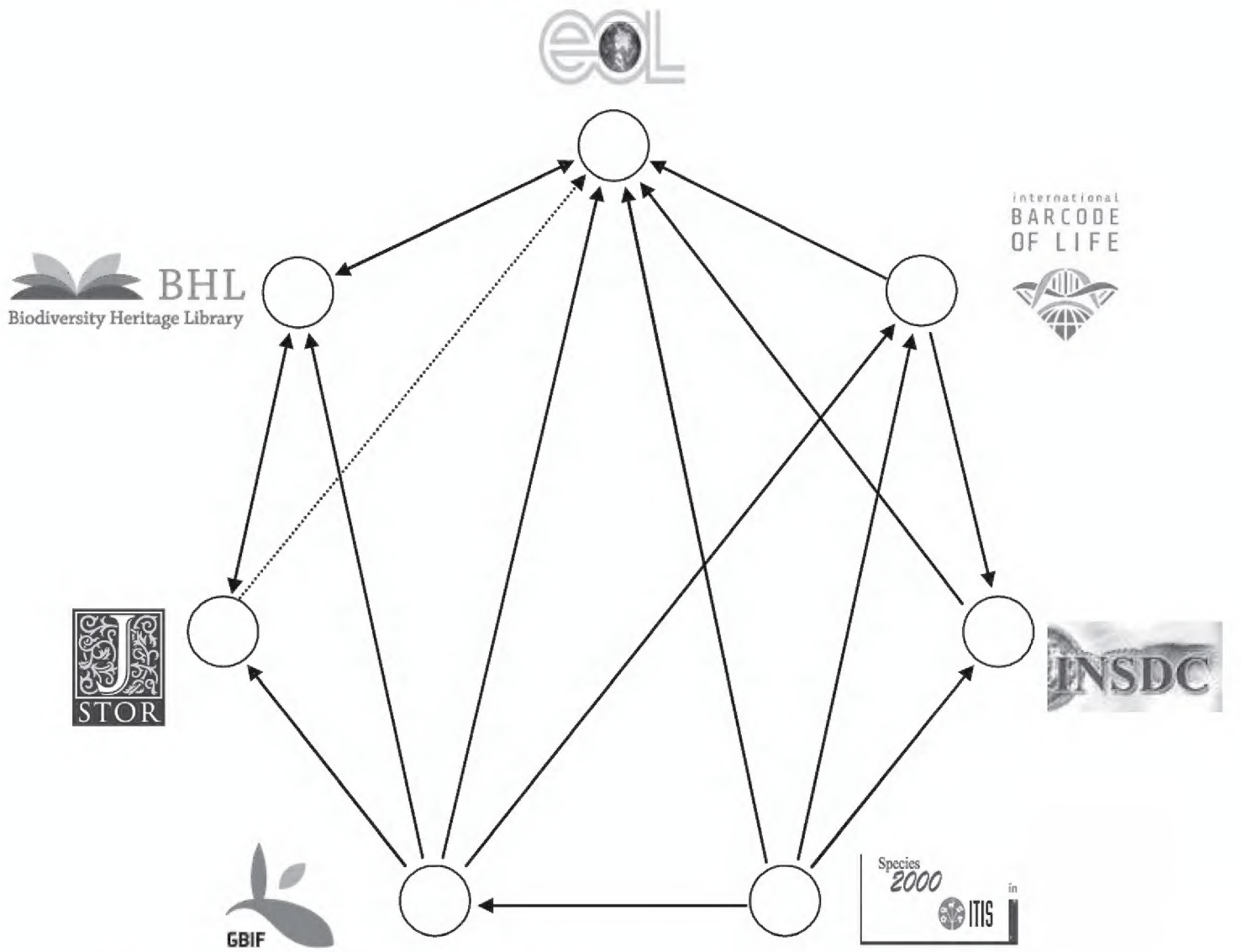

Figure 1. Biodiversity megascience platforms – cross-linkages and data exchange.

ter of Marine Species (WoRMS; http://www.marinespecies.org/), Species Fungorum (http://www.speciesfungorum.org/), and FishBase (http://www.fishbase.org/).

Names data, taxonomy, and classifications are of essential interests for all biodiversity platforms. Thus the comprehensive and reliable species databases offered by CoL form one of the multiple taxonomic backbones of EOL, GBIF, iBOL, BHL, and the INSDC data platforms.

Concerning taxonomic names and classifications, the data flows will be even more complicated in the future because there are overlapping and competing name thesauri for taxonomic and biological groups worldwide. As an example: Lichen names and synonym data are actually being collected by three different major sites (Index Fungorum/ Species Fungorum; http://www.indexfungorum.org, LIAS names, and MycoBank), and are either directly forwarded to several megascience platforms, or indirectly via CoL.

Another type of data flow starts with the occurrence data harvested by the megascience platform GBIF. Several initiatives or projects like EDIT and BioCASE established data flow structures with mirrors of the GBIF index database. Based on these cache databases, they forward large amounts of GBIF occurrence data to various thematic search portals (http://search.biocase.org/; http://search.biocase.org/edit/; Holetschek et al. 2009).

Data harvesting, data exchange, and data quality

Different data harvesting strategies are required (a) for the initial content building from facts not yet available in aggregated form, and (b) for harvesting data that are already aggregated and available as databases, digital publications. In the latter case this may be organized as a unidirectional, perhaps hierarchical data flow, or as reciprocal exchange (partial or full data replication).

In both cases, the goal of megascience platforms is to attract data from a large number of potential provider groups, researchers and research groups, citizen scientists, and established infrastructure and science institutions. With regard to the data domains in focus of JSTOR Plant Science and BHL, institutions are the main data providers, whereas INSDC attract individual researchers and EOL – at least – intends to attract individual researchers and 'citizen scientists' to contribute with their data. Currently, however, the majority of data in EOL comes from other databases: Wikipedia, FishBase, Plazi, etc.

GBIF and CoL address large and small data aggregators, both institutional and individual, but not accept single data records from individual scientists. They require a certain level of aggregation and the capacity to follow structured information transfer protocols according to data exchange standards.

All seven platforms have to be attractive for their data provider communities and use easy-to-use upload techniques, modern web presentation, analysis and visualisation techniques and at least have started the implementation of download options. To facilitate massive collaboration with data providers, data users, and the data exchange between platforms of other data domains, the use of creative commons licenses for data content is urgently recommended (Hagedorn et al. 2011).

EOL was initiated as a funded project and will depend on third-party funds for continued operation. With its strong dependency on biodiversity communities and the activities of individuals and other project content partners, it will always be confronted by new user requirements due to the changing internet world and the rapid enhancement of web technologies. EOL relies mainly on the aggregation and harvesting of external content and uses established web technologies and community solutions to mobilise and cache data. Active input by users is guided via community user interfaces (e.g., until 2010 through so called LifeDesks, now by endorsing ViBRANT scratchpads).

With the growth of content and the rapid enhancement of web technologies, new technical challenges will have to be met to keep large amounts of data manageable and available. Thus the analysis options of the content data for scientific purposes actually are not (yet) in the focus of this platform.

The Wikipedia platform (as well as the associated Wikispecies) goes a citizen science driven and interactive way to mobilise species-related description data and images and provide them to public. Wikispecies currently comprises more than 343,862 content pages (mostly taxon pages, https://species.wikimedia.org/wiki/Special:Statistics), the contents of which is limited to nomenclature, taxonomic hierarchy, or names in various languages. The English Wikipedia contains approximately 213,661 taxon pag-

es (http://toolserver.org/~jarry/templatecount/index.php?lang=en&name=Template%3ATaxobox#bottom), most of which with substantial content.

INSDC is the only platform which has an explicit mandate from the scientific community to harvest and present data. This is achieved through alliances with publishers. Today, the editorial rules of most journals consider INSDC deposition of nucleotide or protein sequences and the citation of the resulting INSDC accession numbers as mandatory, a practice which "arose not passively, but through the efforts of INSDC member institutions and other proponents of open data sharing" (Cochrane et al. 2010). The technical mechanism of the data exchange in the INSDC consortium (with regard to nucleic acid sequence data submission and provision) is the pooling of the original data into one joint data management system, managing this newly established system at one institution and mirroring the database to the consortial partners. iBOL is using the INSDC consortial infrastructure for data archiving.

The large number of providers for occurrence data (from the monitoring community as well as the natural history collection community) and the large amount of data packages which are regularly updated determine the harvesting strategy of the GBIF network. It was originally planned for continuous connectivity and distributed queries, but the technical limitations were difficult to master. GBIF therefore now uses harvesting of a limited set of data instead (called 'indexing'), such that the index is centrally maintained and can be directly queried. With the new GBIF integrated publishing toolkit (IPT) GBIF has been able to support a much wider range of content providers with less technical expertise. The updating of the harvested data may occur at short intervals, or only when a provider publishes a new version. In that way, they underline the decentralized approach of the network with independent data holders or publishers and a mediating role of the national GBIF participant nodes. The new harvesting network of CoL follows a similar strategy.

Data curation and quality control of harvested data is a main issue for all megascience platforms (e.g., Costello et al. 2012). All have to consider quality (in the sense of Chapman 2005) of the original data and address the life cycle of data. They do it in different ways:

GBIF, iBOL, JSTOR Plant Sciences, and probably INSDC, work to establish feedback mechanism to their primary data providers to improve quality of data. GBIF and CoL are planning to realise technical workflows to obtain high-quality data from primary sites by dynamic periodic and event-based data harvesting. Thus, they are likely to provide relatively up-to-date data, as far as the connected primary sites are maintained by domain experts. Platforms like iBOL rely on the direct input and curation efforts of the contributing scientific community and single researchers to ensure and improve the quality of data – similar as INSDC does. Besides relying on the quality of the harvested data from large content partners, EOL has established an own system of single EOL curators, who are expected to improve the harvested EOL content. There is, however, no regular feedback option to the primary data providers.

In addition, copies of harvested data occur which might be harvested again by EOL (or other megascience platforms and thematically focussed portals). Thus, it can

happen that the secondary information becomes ranked higher in internet searches than the original, well-curated information from the primary information site. Information duplication of this kind is most easily visible with Latin taxon names. For instance, a Google search of "*Rimularia exigua*", a hitherto extremely rarely collected crustose lichen from Australia, only having been treated in the context of one monograph and occurring in only one primary species checklist, results in 330 hits, nearly all from secondary and tertiary data harvesters and portals like Cybertruffle (http://www.cybertruffle.org.uk) and SinBiota 2.0 (http://sinbiota.biota.org.br) which spread names data obtained, e.g., from CoL. Unfortunately, not only correct names are disseminated but also misspelled or otherwise erroneous names, even if they are corrected already at a primary information site.

Benefits for data producers, primary data providers and data consumers

Data producers and primary data providers are individuals or organizations that contribute with their data to the content of megascience platforms. They may profit in decidedly different ways from such an activity. The member institutions of JSTOR Plant Science are paid for their digitalisation efforts and contribution to the initiative by the A. Mellon foundation. With regard to GBIF, data providers directly profit from an established data pipeline that allows publishing data sets by using the integrated IPT publishing toolkit as recommended by the GBIF secretariat. In that context, the source data are getting processed and published in standard-compliant Darwin Core Archive (DwC-A) and Ecological Modeling Language (EML v2.1.1) formats (http://www.gbif.org/informatics/infrastructure/publishing/). Various feed-back mechanisms at the GBIF central node support quality control at the primary data site.

The easy access to useful and reliable high-quality data for open and free "data-driven" research purposes (with the aim to publish in high-ranked scientific journals) may be primarily of interest to the platform users and consumers, but not necessarily to the operators and content providers. The content maintenance of a scientific data platform therefore has to be considered as a valuable achievement of the data generators (and maintainers) *per se*. Recently, 'data publishing' through scientific information portals is combined with new kinds of mechanisms to provide additional incentives to data owners that provide their original data to others. The so-called 'data papers', currently promoted by GBIF and EOL community members and publishers like Pensoft (Chavan and Penev 2011), are suggested as an option to form a link between biodiversity data publishing via megascience platforms or portals and the scholarly publishing in peer-reviewed journals with DOI assignment and provision of impact factors. The process of data-paper-publishing uses a common GBIF/Pensoft workflow of data publishing and automated generation of data paper manuscripts using the GBIF integrated publishing toolkit, followed by the editorial workflow via the Pensoft online editorial system and resulting in a regular scholar

publication in online publication like the 'Biodiversity Data Journal' (http://www.biomedcentral.com/1471-2105/12/S15/S2) and MycoKeys.

Reliable and quality-controlled data are a prime interest of data consumers. The data publishing mechanism in the context of INSDC is the best example for that. It requires the active submission of the respective data sets by individuals or organisations which receive an INSDC accession number for every submitted nucleic or amino acid sequence. This identifier is requested by peer-reviewed journals for submission of manuscripts and allows for the backtracking of information to the data producer.

A similar solution is presently being established for the improvement of data content of fungal names thesauri which – regarding the data flow – will secondarily positively influence CoL data. A group of mycologists and database operators gained influence on the fungal scientific community and achieved that the new ICN code (ratified in Melbourne 2011) dictated, that, as of 1 January 2013, each new fungal name must be registered in a recognized repository prior to publication (Norvell 2011, Norvell and Redhead 2012). From a technical point of view, such obligations are probably unnecessary. It seems to make more sense to realise technical solutions for harvesting this type of data from open access (and access-limited) journals, all by now being available in digital form. To do this effectively, markup standards for scientific publishing should be developed, a topic presently dealt with by pro-iBiosphere.

Primary data providers also profit to some degree from seed money projects being funded by platform initiatives and consortia like GBIF, EOL, and CoL. At least, during the first years, iBOL proved to be an excellent opportunity for natural history collections to receive free DNA barcoding data of specimens in their own collections.

Primary data providers usually are also users of their own data and profit from various kinds of analysis options. As data are generally openly accessible (except those in JSTOR; see above), analysis of own data against a wider data background has become a standard use case. Most published phylogenies are based on nucleic acid sequence data of the data producer (or primary provider) combined with otherwise published background sequence data. The situation is similar for occurrence data, where freely available bioinformatics and biodiversity informatics tools for data analysis (INSDC, GBIF, iBOL, and BHL) and visualisation (GBIF, JSTOR, BHL, and EOL) enlarge benefit for platform users.

The benefit for scientists mainly depends on the amount and quality of openly and freely available information. Established megascience information platforms with a history of more than ten years like INSDC already comprise a considerable number of records. However, due to missing or insufficient data curation services by INSDC, insufficient mechanisms to improve and enrich previously submitted (meta-)data, uncritical use of INSDC cannot be recommended. For that reason, a considerable number of thematically focused secondary data platforms have evolved, providing quality-controlled data. In the context of nucleic acid sequence data especially valuable examples are the 'ITS2 Database' at Würzburg University, Germany (http://its2.bioapps.biozentrum.uni-wuerzburg.de), several RNA databases (e.g., http://www.bioexplorer.

net/Databases/RNA_Databases/), or, as an example of a full genome sequence database, the *Saccharomyces* genome database (http://www.yeastgenome.org).

In some cases, the quality of a data may also decrease with time. For instance, data being linked with taxonomic names may degenerate, as taxonomic opinions and phylogenetic concepts are not stable over time. The reasons for this are the discovery of new taxa, the reappraisal of old or discovery of new phenetic traits or of additional gene markers, or the application of improved data analysis algorithms. It entails that under insufficient and inadequate data curation conditions that insufficiently provide for data updates from the original data sources, even well-established megascience platforms are liable to become outdated sooner or later. With regard to taxonomic and nomenclature data flow mechanisms, two major preconditions need to be considered. Firstly, that external taxonomy sources, providing synonymy and classification, are up-to-date and second, that feed-back mechanisms between data sources and platforms need to provide mechanisms for correcting recognized inconsistencies. Both issues are presently not satisfactorily realized even for the oldest megascience platform INSDC, despite the fact that this platform has probably the strongest profile of all established biodiversity information platforms under discussion.

Discussion

In an era of data-driven research and open science (Krotoski 2012), biodiversity data platforms are facing a number of challenges. Perhaps the most important issue is the question of sustainability in data curation and software development. Data curation is a complex task that involves both primary data producers or providers and platforms which integrate such data. Although a primary responsibility for correctness lies with the primary data producers or providers, the platform has a responsibility to monitor the data quality and the frequency of updates from the data sources. A considerable part of quality control concerns the necessity of a data integration workflow, which typically exposes data quality issues, that where difficult to detect, while the data were curated in isolation. Beyond that, many platforms invest into purpose-built quality control tools, drawing on the development, computing, and data source integration power of the platform. Since the platform is often attracting a much larger number of users than the primary data source (should it be online), much feedback and annotation activity is likely to occur on the platform. Both, the platform workflow or tools-supplied and user-supplied feedback must be efficiently communicated to the primary data sources.

Amount and granularity of the primary data sources that are harvested or integrated into the platforms can range from huge databases to individual contributions both with elementary or rather detailed information. Although the various platforms have a different focus, in fact all have to support a wide spectrum of granularity from individuals to institutions. Because individuals typically have rather different means as well as motivations to curate a dataset than institutions, this further complicates quality control, annotation and feedback workflow. Presently, megascience platforms

rarely include the publishing level, which can be seen as a granularity gap between individual contributions (by direct editing) and data flow from private or institutional databases. New efforts (e.g., within the pro-iBiosphere project) explore the necessary collaboration infrastructure for a biodiversity 'Knowledge Organisation System' that bridges existing gaps between scientific publishing (journal articles as well as flora/fauna monographs) and megascience data platforms. To enable integration, structuring, quality control, feedback mechanism, attractive data retrieval and other sophisticated services (e.g., Hill et al. 2010), or even the realisation of virtual research environments, platforms need to invest into man person-years of software development work. A major problem with respect to the present dynamic world of a global information system is that software needs constant investment in maintenance and development simply to keep up with ongoing feature development and security fixes of the basic tools as well as software interfaces of partners.

Furthermore, the number of platforms with thematic but global focus in biology and environmental sciences is increasing. In the field of biodiversity they are often backboned by automatically generated template web pages filed according to taxon names. The temptation to fill these auto-generated pages with existing name lists and classification structures is evident and somehow understandable as it serves the desire to become globally relevant. The hope that such templates will be supplied with content by scientific community members, however, is rarely fulfilled.

The relation between megascience biodiversity information platforms and smaller, more focussed data providers is and will remain a complex one. Simplifying it by shifting all responsibility and ownership of data to a central institution or data node may, however, not be the right path into the future. While focussed central platforms can become a service to stakeholders, all-encompassing platforms are likely to satisfy only a limited number of use-cases. As a result, stakeholders still would require independent systems, leading in the end to lower total efficiency. We therefore believe that sharing responsibility and funding opportunities is the right path into the future. For the content partners of megascience biodiversity information platforms, it is most likely to be beneficial, if they operate their own original or primary databases under their own responsibility at an institution. In the long term that means – from the view of the megascience platforms – a decentralised approach should be realised. In that way, data sustainability and quality seems to be best ensured. The technical support for primary-content databases should be guaranteed by commitments of the institutions which hosts or own the databases. Also at that level of a decentralised biodiversity data network data architecture and IT infrastructure have to be continuously adapted to the changing requirements. At the same time, the infrastructure of the megascience platforms also depend on institutional or other reliable and permanent funding, as the technical and content data management of the platforms themselves will always remain a challenging task.

Due to the steadily increasing number of scientists from countries all over the world being involved in higher level biodiversity and environmental science projects, it is clear that certain architectures and mechanisms of data storage, transfer and provi-

sion will be recognized as obsolete. They are symptomatic of a past unilateral world. The megascience platforms discussed here, have to attract both, new primary-content partners by offering added values to them as well as new technical partners, e.g. as consortial members of equal rank. To be able to replicate information with primary-content partners, it will be necessary to implement technical interfaces that better support data exchange standards. In recent years with the rise of new user interface concepts, the mode of presentation needs to be adapted to changes in the device technologies (gestures and touch modes). Alleged limitations of database and data transfer technologies are sometimes used as an alibi to replace federated structures of distributed responsibility and ownership with central and often 'monopolistic' structures. However, centralised power always includes the temptation of abuse, be it to dictate prices (as seen in some major commercial scientific publishers), or be it to monopolize the use of data for research, trying to secure future research grants at the expense of excluding competing researchers (which may have a different research agenda, perspective, or insight).

Both single and distributed ownership of primary data can lead to monopolies or single-points of failure (for all or parts of the data). It is not uncommon that valuable data sources are either lost or that the owners decide to no longer share them. Long-term preservation and open access to scientific data is a prime value in science. Both a system of a single platform with a single data store, and a system where a large number of stakeholders could arbitrarily decide that it is no longer financially feasible or perhaps desirable to them to provide their data to the scientific community, does not fulfil this requirement. The solution would have to provide for a large number of duplicated storage of data, the use of which is at least as uninhibited as the use of books. Achieving this is (a) a technical problem in finding the right technologies to replicate large volumes of data, (b) a social problem in documenting and understanding the difference between primary holders that frequently update their data versus static copies that have been created for particular uses and which may become outdated, and (c) a legal problem, in providing sufficient rights over the copied data. Scientific knowledge becomes more valuable to society, the more it is shared. The scientific world must therefore take care that the principles of openness and sharing that have successfully governed science for centuries are not lost in the new age of digital scientific data. Sharing has to be open and permissive, following the principles of Open Science, Open Source and Open Data (Molloy 2011).

The megascience platforms discussed here already have to face complementary or alternative structures (e.g., EOL China, http://www.eolchina.org/; Species2000 China Node, http://www.sp2000.cn/joaen/; BHL China, http://www.bhl-china.org/cms/). Global platforms will probably still dominate in the near future and guide mainstream activities, but they will not be able to claim an exclusive status. They are driven by modern information technologies and have to support approaches for decentralized and 'intelligent' network structures with flexible data nodes. In this context, efforts of multilinguality and internationalisation should also be prioritized. Despite English being de facto the lingua franca of natural sciences, IT technologies will increasingly allow to (automatically) generate multilingual presentations to include users from countries outside the space of world-dominating languages.

Acknowledgements

The work was supported in part by the Federal Ministry of Education and Research, Germany (BMBF) with the project 01 LI 1001 B 'GBIF-D' and by the German Research Foundation (DFG) with the LIS infrastructure program grants INST 747/1-1, RA 731/11-2, and TR 290/5-1. Support was also granted by the European Union's 7th Framework Programme (FP7/2007-2013) with the projects 4D4Life (grant agreement №238988), ViBRANT (grant agreement №261532) and pro-iBiosphere (grant agreement №312848).

References

Chapman AD (2005) Principles of Data Quality. Copenhagen: Global Biodiversity Information Facility, 58 pp.

Chavan V, Penev L (2011) The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics 12 (suppl. 15): S2. http://www.biomedcentral.com/1471-2105/12/S15/S2, doi: 10.1186/1471-2105-12-S15-S2

Cochrane G Karsch-Mizrachi I, Nakamura Y (2010) On behalf of the International Nucleotide Sequence Database Collaboration. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Research 39: D15–D18. doi: 10.1093/nar/gkq1150

Costello MJ, Michener WK, Gahegan M, Zhi-Qiang Zhang Z-Q, Bourne P, Chavan V (2012) Quality assurance and intellectual property rights in advancing biodiversity data publications ver. 1.0, Copenhagen: Global Biodiversity Information Facility, 33 pp. http://links.gbif.org/qa_ipr_advancing_biodiversity_data_publishing_en_v1

Hagedorn G, Mietchen D, Morris RA, Agosti D, Penev L, Berendsohn WG, Hobern D (2011) Creative Commons licenses and the non-commercial condition: Implications for the re-use of biodiversity information. In: Smith V, Penev L (Eds) e-Infrastructures for data publishing in biodiversity science. ZooKeys 150: 127–149. doi: 10.3897/zookeys.150.2189

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society London B, 270: 313–321. doi: 10.1098/rspb.2002.2218

Hill AW, Otegui J, Ariño AH, Guralnick RP (2010) GBIF Position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF Network, version 1.0., Copenhagen: Global Biodiversity Information Facility, 25 pp. http://www.gbif.org

Holetschek J, Kelbert P, Müller A, Ciardelli P, Güntsch A, Berendsohn WG (2009) International networking of large amounts of primary biodiversity data. – In: Fischer S, Maehle E, Reischuk R (Eds) Informatik 2009 – Im Focus das Leben. – Beiträge der 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI). GI-Edition: Lecture Notes in Informatics (LNI), Proceedings 154: 23, 552–564. http://subs.emis.de/LNI/Proceedings/Proceedings154/gi-proc-154-8.pdf

Kasperek G (2010) Overview of historical biology literature digitisation projects – compiled by vifabio, the Virtual Library of Biology. Newsletter BHL-Europe, No. 5/6: 11–15. http://www.vifabio.de/about/files/pub/kasperek_2010_BHL-Europe_Newsletter_5-6.pdf

Krotoski AK (2012) Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise. Insights: the UKSG journal, 25(1): 28–32.

Molloy JC (2011) The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biol 9(12): e1001195. doi: 10.1371/journal.pbio.1001195

Norvell LL (2011) Fungal Nomenclature. 1. Melbourne approves a new Code. Mycotaxon 116: 481–490. doi: 10.5248/116.481

Norvell LL, Redhead SA (2012) Stop Press! Registries of names and the new Code. – sIMA Fungus 2(1): 2. http://www.imafungus.org/Issue/31/03.pdf

Rambold G (lead editor) (2012) A Global Information System for Lichenized and Non-Lichenized Ascomycetes. Col – Global Species Database for Lichens. http://www.catalogueoflife.org/col/details/database/id/

Triebel D, Neubacher D, Weiss M, Heindl-Tenhunen B, Nash TH III, Rambold G (2010) Integrated biodiversity data networks for lichenology – data flows and challenges. – In: Nash TH III, Geiser L, McCune B, Triebel D, Tomescu AMF, Sanders WB (Eds) Biology of lichens – symbiosis, ecology, environmental monitoring, systematics and cyber applications. – Biblioth. Lichenol. 105: 47–56.

Internet resources

Atlas of Living Australia (ALA) – http://www.ala.org.au/

BioCASE Search Portal – http://search.biocase.org/

Biodiversity Heritage Library (BHL) – http://www.biodiversitylibrary.org; http://biodiv-lib.wikispaces.com

Biodiversity Heritage Library China (BHL China) – http://www.bhl-china.org

Bioexplorer – http://www.bioexplorer.net

BioMed Central – http://www.biomedcentral.com

Biowikifarm – http://biowikifarm.net

BOLDSYSTEMS – http://www.boldsystems.org; http://www.barcodinglife.com

Catalogue of Life (COL) – http://www.catalogueoflife.org; http://www.catalogueoflife.org/colwebsite/content/contributors/

Cybertruffle – http://www.cybertruffle.org.uk

Distributed Dynamic Diversity Databases for Life (4D4Life) – http://www.4d4life.eu/

DNA Data Bank of Japan (DDBJ), Mishima, Japan – http://www.ddbj.nig.ac.jp

EDIT Search Portal – http://search.biocase.org/edit/

EMBL-Bank, European Nucleotide Archive, Cambridge, UK – http://www.ebi.ac.uk/embl/

Encyclopedia of Life (EOL) – http://eol.org

Encyclopedia of Life China (EOL China) – http://eolchina.org

European Distributed Institute of Taxonomy (EDIT) – http://www.e-taxonomy.eu/

FishBase – http://www.fishbase.org/

Fungal Names Registration – http://www.fungalinfo.net

Global Biodiversity Information Facility (GBIF) – http://www.gbif.org

Global Names Architecture (GNA) – http://www.globalnames.org/
GenBank, NCBI, Bethesda, MD, USA – http://www.ncbi.nlm.nih.gov/genbank/
Global Plants Initiative (GPI) – http://gpi.myspecies.info/; http://plants.jstor.org/action/community/
Index Fungorum – http://www.indexfungorum.org
International Barcode of Life (iBOL) – http://www.barcodinglife.com
International Nucleotide Sequence Database Collaboration (INSDC) – http://www.insdc.org
ITS2 Database – http://its2.bioapps.biozentrum.uni-wuerzburg.de
JSTOR – http://www.jstor.org
JSTOR Plant Science – http://plants.jstor.org
LIAS light – http://liaslight.lias.net
LIAS names – http://liasnames.lias.net/
LifeWatch – http://www.lifewatch.eu
Map of Life (MOL) – http://www.mappinglife.org
MycoBank – http://www.mycobank.org/
pro-iBiosphere – Coordination and policy development in preparation for a European Open Biodiversity Knowledge Management System, addressing Acquisition, Curation, Synthesis, Interoperability and Dissemination – http://www.pro-ibiosphere.eu/
Saccharomyces Genome Database (SGD) – http://www.yeastgenome.org
Scratchpads biodiversity online – http://scratchpads.eu/
SinBiota 2.0 – http://sinbiota.biota.org.br
Species Fungorum – http://www.speciesfungorum.org/
Species 2000 China Node – http://www.sp2000.cn
The Mycology Net – http://www.mycology.net
uBio Indexing & Organizing Biological Names – http://names.ubio.org
Virtual Biodiversity Research and Access Network for Taxonomy (ViBRANT) – http://vbrant.eu/
Wikimedia Toolserver – https://toolserver.org
Wikispecies – https://species.wikimedia.org
World Register of Marine Species (WoRMS) – http://www.marinespecies.org/